

# Now, whose schools are really better (or weaker) than Germany's? A multiple testing approach

Christoph Hanck

Universiteit Maastricht

November 2008

# Now, Whose Schools are *Really* Better (or Weaker) than Germany's? A Multiple Testing Approach

Christoph Hanck*

November 2008

**Abstract**

Using PIRLS (Progress in International Reading Literacy Study) data, we investigate which countries' schools can be be classified as significantly better or weaker than Germany's as regards the reading literacy of primary school children. The 'standard' approach is to conduct separate tests for each country relative to the reference country (Germany) and to reject the null of equally good schools for all those countries whose $p$-value satisfies $p_i \leqslant 0.05$. We demonstrate that this approach ignores the multiple testing nature of the problem and thus overstates differences between schooling systems by producing unwarranted rejections of the null. We employ various multiple testing techniques to remedy this problem. The results suggest that the 'standard' approach may overstate the number of significantly different countries by up to 30%.

*Keywords*: PIRLS, Multiple Testing, Multi-Country Comparisons

*JEL classification*: C12, I21

*Universiteit Maastricht, Tongersestraat 53, 6211 LM Maastricht, The Netherlands. Tel. (+31) 43-3883815, c.hanck@ke.unimaas.nl.

# 1 Introduction

Multi-country comparisons of student achievement regularly cause lively debate, both in academic circles and, perhaps even more so, the wider public. In view of below-average results in recent Programme for International Student Assessments (PISA), this seems to be particularly the case in Germany. Germany's place in the international ranking is tracked closely across different editions of the assessment, and comparatively better performances in other exercises, such as PIRLS (Progress in International Reading Literacy Study), are widely acclaimed. Of course, the professional literature (as well as the better media) recognizes that a country being placed before another one in a ranking does not necessarily have a better educational system than worse-placed one. To keep the effort of student assessments manageable, all of these are inevitably based on samples from the countries' student population. Thus, any comparison of any two country must make use of the tools of statistical inference. In particular, it is to be investigated whether differences found between two countries are statistically significant. If the only analysis of interest was one *single* comparison of two countries, this could be done routinely with, say, a suitable $t$-test.

However, large-scale international student assessments typically have several dozens of participating countries. The relevant issue then becomes to test whether *any* of $n$ countries' schools is better (or weaker) than the reference country of interest. The literature typically investigates this question by conducting separate $t$-tests for each country relative to the reference country, and declares all those countries' schools as significantly different from the reference country's for which the corresponding $p$-value is sufficiently small, say, $p_i \leqslant 0.05$ [e.g., Bos, Hornberg, Arnold, Faust, Fried, Lankes, Schwippert, and Valtin, 2007]. In the sequel, we shall refer to this approach as the 'standard' approach.

Unfortunately, this simple and intuitive way of investigating whether countries', say, reading performances are significantly different from each other is problematic from a statistical point of view. Effectively, it ignores the multiple testing nature inherent of the approach. To illustrate the problem, consider the following artificial numerical example. Suppose one has achievement data on a panel of, say, $n = 20$ countries (plus one reference country). Also assume for simplicity that the the countries are statistically independent and that all countries' performances are identical.[1] When conducting tests on each country at the $\alpha = 0.05$ level, one might casually expect the probability to erroneously find evidence in favor of significant differences in at most one case to equal 5%, because $1/20 = 0.05$. However, the event of a rejection is a Bernoulli random variable with "success" probability 0.05. Hence, $P_k$, the probability of finding $k$ rejections in $n$ tests, is the probability mass function of a Binomial random variable,

$$P_k = \left( \begin{array}{c} n \\ k \end{array} \right) \alpha^k (1-\alpha)^{n-k}.$$

Therefore, the probability of (at least) one erroneous rejection, also known as the Familywise Error Rate[2] ($FWER$), equals

$$\begin{aligned} P_{k \geqslant 1} &= \sum_{j=1}^{20} \left( \begin{array}{c} 20 \\ j \end{array} \right) 0.05^j (1-0.05)^{20-j} \\ &= 1 - P_0 \\ &= 1 - \left( \begin{array}{c} 20 \\ 0 \end{array} \right) 0.05^0 (1-0.05)^{20} \\ &= 0.6415. \end{aligned}$$

Even if all countries have identically good schools, one will falsely find some evidence of

[1] This assumption is only made to justify the following calculation. It is not needed for any of the procedures we shall employ later.

[2] Let $P$ be the true data generating mechanism and $I_0(P) \subset \{1, \ldots, n\}$ the units for which corresponding null hypothesis $H_i$ is true. A precise definition is then given by

$$FWER_P = \Pr_P\{\text{Reject at least one } H_i : i \in I_0(P)\}$$

More generally, the *j-FWER* is defined as $P_{k \geqslant j}$, the probability of $j$ or more false rejections.

differences with a rather high probability. Of course, the problem only worsens if one adds more units to the panel. If one has data on a broader set of 100 countries, the corresponding probability equals 0.9941. That is, one is then practically bound to declare at least one—and potentially quite a few more—countries' schools as significantly different from the average even if all are equal.

This so-called "multiplicity" problem, while not widely recognized in the broader econometrics literature [Savin, 1984], has of course been realized long ago in the statistics literature [see Lehmann and Romano, 2005]. Several solutions to controlling the $FWER$ at some specified level $\alpha$ have been suggested. Among the most popular are the Bonferroni and the the Holm [1979] procedure. These procedures have however been less successful in applications because ensuring $FWER \leqslant \alpha$ typically comes at the price of reducing the ability to identify false hypotheses. That is, the procedures are conservative or have low "power."[3] Hence, often quite reasonably, researchers have tended to ignore the issue of multiplicity.

There has been substantial research on improving the ability of multiple testing approaches to detect false hypotheses while still controlling the $FWER$. Romano and Wolf [2005] put forward a bootstrap scheme that exploits the dependence structure of the statistics in order to improve the power of the multiple test. Hommel [1988], in turn, works with a computationally less demanding $p$-value combination technique. Benjamini and Hochberg [1995] suggest a procedure that is likely to detect more false hypotheses in particular in situations with a large $n$.

The present study uses data from the 2006 edition of the PIRLS assessment—better known as IGLU in Germany—to investigate the effect of multiplicity on the classification of countries' schools into those better or weaker than Germany's. We analyze which countries

[3]For a discussion of "power" in a multiple testing framework see Romano and Wolf [2005], Sec. 2.2.

have better or weaker schools as regards the reading literacy of 4th grade students. Our main finding is that not controlling for multiplicity via suitable multiple testing techniques overstates the number of significantly different countries by up to 30%.

The next section summarizes the multiple testing procedures used in the present study. Section 3 provides some background on the PIRLS study. Section 4 presents the empirical results, while the last section summarizes and provides an outlook for possible further research.

# 2 Multiple Testing Procedures

We now briefly outline the multiple testing procedures used here. For a full discussion of the properties of the procedures, the reader is referred to the original contributions. Also, Dudoit and van der Laan [2007] and Romano and Wolf [2008] provide recent surveys of the literature.

## 2.1 Classical *FWER*-controlling techniques

Probably the most widely used techniques to control the *FWER* are the Bonferroni and the Holm [1979] procedures. Recall that the former rejects the null hypothesis $H_i$ if the $p$-value $p_i$ corresponding to the test statistic $\hat{\tau}_i$ satisfies $p_i \leqslant \alpha/n$. The Holm [1979] procedure first sorts the $p$-values from smallest to largest, $p_{(1)} \leqslant \ldots \leqslant p_{(n)}$. Relabel the hypotheses accordingly as $H_{(k)}$. Then, reject $H_{(k)}$ at level $\alpha$ if

$$p_{(j)} \leqslant \alpha/(n-j+1) \quad \text{for all} \;\; j \in \mathbb{N}_k,$$

with $j \in \mathbb{N}_k$ shorthand for $j = 1, \ldots, k$. The cutoff value for the first hypothesis is identical for both methods, but unlike the Bonferroni method, the Holm [1979] procedure uses gradually less challenging criteria for $H_{(2)}, \ldots, H_{(n)}$.

TABLE I—HOMMEL'S [1988] PROCEDURE FOR $n = 3$

| | | |
|---|---|---|
| $i = 1$: | $k = 1$ | $p_{(3-1+1)} = p_{(n)} > \alpha$ |
| $i = 2$: | $k = 1$ | $p_{(3-2+1)} = p_{(n-1)} > \alpha/2$ |
| | $k = 2$ | $p_{(3-2+2)} = p_{(n)} > \alpha$ |
| $i = 3$: | $k = 1$ | $p_{(3-3+1)} = p_{(1)} > \alpha/3$ |
| | $k = 2$ | $p_{(3-3+2)} = p_{(n-1)} > 2\alpha/3$ |
| | $k = 3$ | $p_{(3-3+3)} = p_{(n)} > \alpha$ |

## 2.2 Hommel's Procedure

Hommel [1988] suggests the following procedure to control the *FWER*.

HOMMEL'S PROCEDURE

A. Compute

$$j = \max\{i \in \mathbb{N}_n : p_{(n-i+k)} > k\alpha/i \text{ for } k \in \mathbb{N}_i\}. \tag{1}$$

B1. If the maximum does not exist, reject all $H_i$ ($i \in \mathbb{N}_n$).

B2. If the maximum exists, reject all $H_i$ with $p_i \leqslant \alpha/j$.

For concreteness, consider an illustrative example where $n = 3$. We find $j$ as the largest $i$ such that all adjacent conditions in Table I hold. If $j$ does not exist, then even $p_{(n)} \leqslant \alpha$, so that we can 'safely' reject all hypotheses. If the $p$-values are given by, say, $3\alpha/5$, $2\alpha$, $\alpha/5$, then $j = 2$ such that we only reject $H_3$.

Graphically, the procedure works as sketched in Figure I. We depict $n = 5$ sorted $p$-values and take $\alpha = 0.05$. In this case, $j = 2$ because, starting from the left, the second-to-last of the blue (darker) lines is the first one such that all corresponding sorted $p$-values are above that line. Hence, we reject all those $H_i$ for which $p_i \leqslant \alpha/2$. That is, the first three hypotheses.

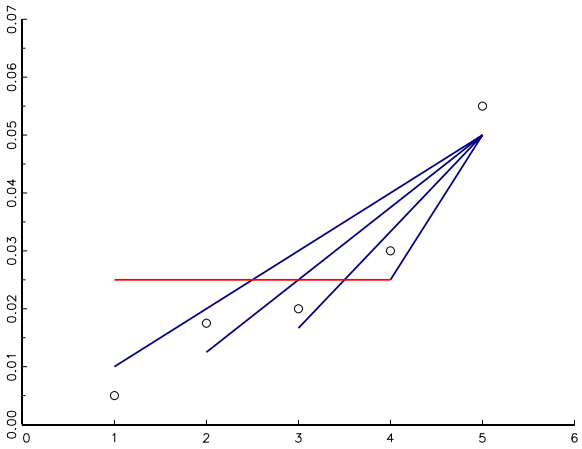



Figure I—A Graphical Illustration of Hommel's Procedure

Originally, Hommel's Procedure was only known to control the *FWER* under independence, an overly strong assumption in our setup. Indeed, student achievement is likely to be affected by background variables such as the extent to which learning is valued in a society. Since that valuation is typically more prevalent in certain geographically related groups of countries, achievement data will not be independent from one country to the next. Fortunately, Sarkar [1998] shows that the assumption of independence is not necessary and can, in fact, be weakened substantially. The following is adapted from Sarkar [1998, Prop. 3.1][4]

Proposition 1.

*If the test statistics for testing the* $H_i$, $i \in \mathbb{N}_n$, *are* multivariate totally positive of order 2 *(*$MTP_2$*), then, for* $j$ *from* (1),

$$\mathrm{P}_{H_0}(\exists\, i \in \mathbb{N}_n : p_i \leqslant \alpha/j) = \mathrm{P}_{H_0}(\textit{Hommel rejects for some } i) \leqslant \alpha,$$

*where* $\mathrm{P}_{H_0}$ *denotes the probability under the null hypothesis.*

[4]In fact, Hommel [1988] proves that his procedure controls the *FWER* if the intersection test by Simes [1986] is a level $\alpha$ test. Sarkar [1998], in turn, proves that Simes' test is a level $\alpha$ test under the $\mathrm{MTP}_2$ condition.

A vector of random variables $T = (T_1, \ldots, T_n)'$ is said to be $\text{MTP}_2$ if its joint density $f$ satisfies

$$f\big(\min(T_1, U_1), \ldots, \min(T_n, U_n)\big) \cdot f\big(\max(T_1, U_1), \ldots, \max(T_n, U_n)\big) \geqslant f(T_1, \ldots, T_n) \cdot f(U_1, \ldots, U_n),$$

for any two points $(T_1, \ldots, T_n)$ and $(U_1, \ldots, U_n)$. The $\text{MTP}_2$ class is rather large, including the multivariate normal with nonnegative correlations, the absolute-valued multivariate normal with some specific covariance structures, multivariate gamma, absolute-valued central multivariate $t$, and central multivariate $F$ distributions. Sarkar [1998] further verifies that even the $\text{MTP}_2$ condition of Proposition 1 is not necessary.

## 2.3 Benjamini and Hochberg [1995]

When the number of multiple tests $n$ is large, control of the *FWER* is often an overly strict criterion, as ensuring a low probability of only one false rejection then comes at the price of low power of the procedures. Also, one might be willing to tolerate more than one false rejection if there are a larger number of total rejections. Put differently, one might be willing to tolerate a small share of false rejections out of the total rejections. To that end, Benjamini and Hochberg [1995] suggest the "False Discovery Rate" (*FDR*). Let $V_n$ the number of false rejections and $R_n$ the total number of rejections. The *FDR* is then defined as

$$FDR = \text{E}\left[\frac{V_n}{R_n}\mathbb{I}(R_n > 0)\right]$$

A multiple testing method is said to control the *FDR* at level $\gamma$ if $FDR \leqslant \gamma$ for any $P$. Unless all null hypotheses are true, the *FDR* is a more liberal error rate. That is, if a procedure controls the *FWER*, it will also control the *FDR*, but generally not vice versa. (If $V_n > 0$, $FWER = \text{E}[\mathbb{I}(V_n > 0)] \geqslant \text{E}[(V_n/R_n)\mathbb{I}(R_n > 0)] = FDR$, because

$(V_n/R_n) \leqslant 1$.) The Benjamini and Hochberg [1995] is a "stepup" method, which first examines the largest $p$-value, and then proceeds "up" to the more significant hypotheses. It works as follows.

A. Sort the $p$-values from small to large,

$$p_{(1)} \leqslant \cdots \leqslant p_{(n)}.$$

Relabel the hypotheses accordingly as $H_{(k)}$.

B. Choose some (small) $\gamma$.

C. Define

$$j^* = \max\{j \in \mathbb{N}_n : p_{(j)} \leqslant \gamma_j\} \qquad \text{where} \quad \gamma_j = \frac{j}{n}\gamma$$

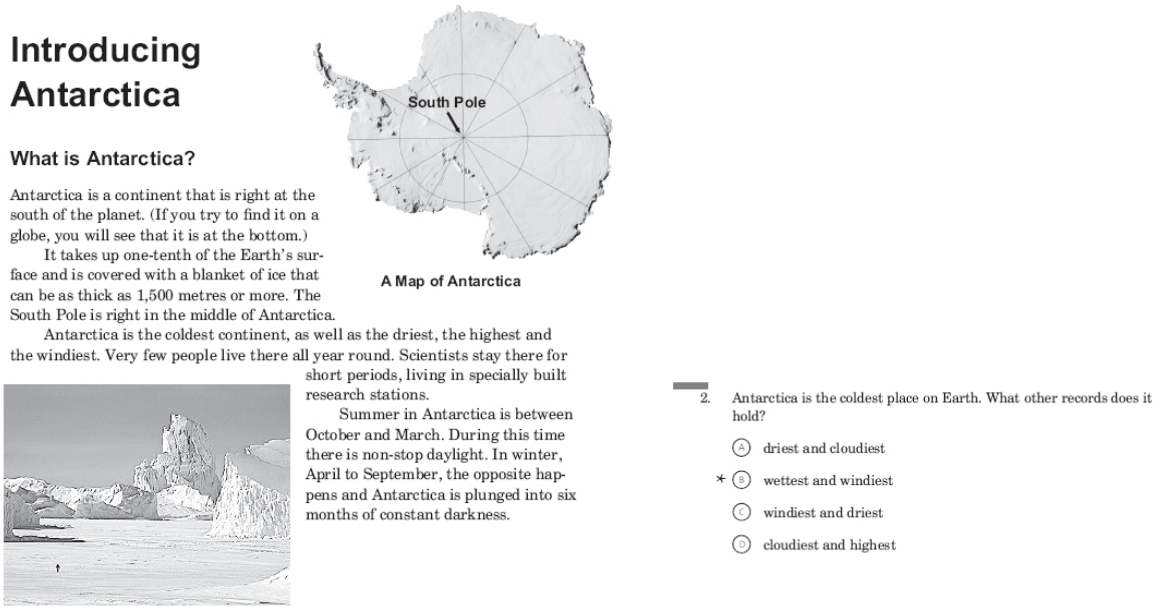
D1. If $j^*$ exists, reject $H_{(1)}, \ldots, H_{(j)^*}$.

D2. If not, reject no hypotheses.

Benjamini and Hochberg [1995] show this procedure to control the *FDR* at $\gamma$ under independence of the test statistics. Importantly, Benjamini and Yekutieli [2001] extend this result and show that this procedure controls the *FDR* under the more general and practically relevant "positive regression dependency on each one from a subset" (PRDS) condition. See Benjamini and Yekutieli [2001] for a precise definition of the PRDS condition, which is somewhat similar to the $\text{MTP}_2$ condition stated above.

# 3 PIRLS

Inaugurated in 2001 and conducted every 5 years, PIRLS—Progress in International Reading Literacy Study—is an assessment of students' reading achievement in the fourth grade (9-10 years old). It aims to monitor international trends in primary school reading achievement [Mullis, Martin, and Foy, 2007]. The 2006 edition of PIRLS was implemented in 40



**Introducing Antarctica**

**What is Antarctica?**

Antarctica is a continent that is right at the south of the planet. (If you try to find it on a globe, you will see that it is at the bottom.)

It takes up one-tenth of the Earth's surface and is covered with a blanket of ice that can be as thick as 1,500 metres or more. The South Pole is right in the middle of Antarctica.



A Map of Antarctica

Antarctica is the coldest continent, as well as the driest, the highest and the windiest. Very few people live there all year round. Scientists stay there for short periods, living in specially built research stations.

Summer in Antarctica is between October and March. During this time there is non-stop daylight. In winter, April to September, the opposite happens and Antarctica is plunged into six months of constant darkness.

2. Antarctica is the coldest place on Earth. What other records does it hold?

A driest and cloudiest

* B wettest and windiest

C windiest and driest

D cloudiest and highest

FIGURE II—A TEXT AND A QUESTION FROM THE PIRLS STUDY

countries, including Belgium with 2 educational systems and Canada with 5 provinces. This yields 45 participants in total—see Figure II for a list of participating countries.[5] Each country selected at least 150 schools, some of which were subsequently excluded from the sample. See Mullis et al. [2007] for a detailed discussion of the reasons.[6]

Figure II depicts an example of a text to be read by students, as well as a corresponding question. (Of course, texts and questionnaires were provided in the students' mother tongues.)

Table II provides some descriptive statistics. It is readily apparent that the countrywise averages have a rather skewed distribution, with many countries exceeding the international average of 500, and correspondingly fewer countries having a substantially lower average score. We further see that there is quite some variation in the number of partici-

[5]Iceland and Norway participated with both the 4th and the 5th grade, which, for the purposes of the present analysis, are treated as two separate units. For brevity, we shall henceforth refer to countries/regions as simply countries.

[6]The data are available at http://www.timss.bc.edu/.

Table II—Descriptive School-Level Statistics.

| Country | Average | Standard Deviation | Number of Schools |
|---|---|---|---|
| Russian Federation. RUS | 569.32 | 67.54 | 232 |
| Hong Kong. HKG | 565.25 | 58.92 | 144 |
| Canada, Alberta. CAB | 560.51 | 67.12 | 150 |
| Canada, British Columbia. CBC | 558.09 | 68.66 | 148 |
| Luxembourg. LUX | 557.09 | 66.73 | 178 |
| Hungary. HUN | 556.41 | 68.19 | 149 |
| Bulgaria. BGR | 555.10 | 81.07 | 143 |
| Singapore. SGP | 552.74 | 78.42 | 178 |
| Italy. ITA | 551.79 | 67.14 | 150 |
| Iceland (5th grade) IS5 | 551.11 | 62.39 | 35 |
| Netherlands. NLD | 550.78 | 51.78 | 139 |
| Denmark. DNK | 549.74 | 69.25 | 145 |
| Sweden. SWE | 549.05 | 63.34 | 147 |
| Belgium (Dutch). BFL | 548.39 | 55.45 | 137 |
| Latvia. LVA | 548.11 | 60.18 | 147 |
| Germany. DEU | 548.07 | 65.20 | 405 |
| Lithuania. LTU | 540.77 | 56.59 | 146 |
| Canada, Ontario. COT | 540.08 | 72.14 | 180 |
| Canada, Nova Scotia. CNS | 537.67 | 76.62 | 201 |
| Austria. AUT | 537.35 | 63.09 | 158 |
| Norway (5th grade). NO5 | 537.31 | 60.62 | 66 |
| United States. USA | 536.84 | 73.67 | 183 |
| England. ENG | 536.71 | 86.61 | 148 |
| Taiwan. TWN | 536.47 | 63.76 | 150 |
| Slovakia. SVK | 535.55 | 74.21 | 167 |
| Canada, Quebec. CQU | 530.35 | 65.59 | 185 |
| Scotland. SCO | 529.88 | 78.93 | 130 |
| New Zealand. NZL | 526.13 | 90.93 | 243 |
| Poland. POL | 524.09 | 74.90 | 148 |
| Slovenia. SVN | 523.06 | 69.75 | 145 |
| France. FRA | 522.84 | 66.74 | 169 |
| Spain. ESP | 516.77 | 69.31 | 152 |
| Israel. ISR | 511.47 | 99.85 | 149 |
| Iceland. ICE | 510.65 | 67.52 | 128 |
| Moldova. MDA | 502.26 | 68.26 | 150 |
| Belgium (French). BFR | 500.43 | 68.62 | 150 |
| Romania. ROM | 500.04 | 88.49 | 146 |
| Norway. NOR | 497.21 | 67.47 | 135 |
| Georgia. GEO | 474.14 | 72.83 | 149 |
| Trinidad and Tobago. TTO | 445.97 | 102.84 | 147 |
| Macedonia. MKD | 442.85 | 100.02 | 147 |
| Iran. IRN | 435.56 | 96.74 | 236 |
| Indonesia. IDN | 405.67 | 79.04 | 168 |
| Qatar. QAT | 354.31 | 96.24 | 119 |
| Kuwait. KWT | 330.56 | 110.30 | 149 |
| Morocco. MAR | 328.04 | 106.66 | 159 |
| South Africa. ZAF | 282.02 | 119.17 | 397 |

pating schools per country, with some countries such as Germany (405) strongly exceeding the target value of 150, and others falling short.

# 4 Results

As argued in the Introduction, the extent to which different educational systems are found to be statistically significantly different may be overstated in the literature, as the employed testing procedures typically do not account for multiplicity. This section presents an application of the above multiple testing procedures to identify those countries that have statistically significantly better primary schools than Germany as regards reading literacy, while controlling for multiplicity.[7]

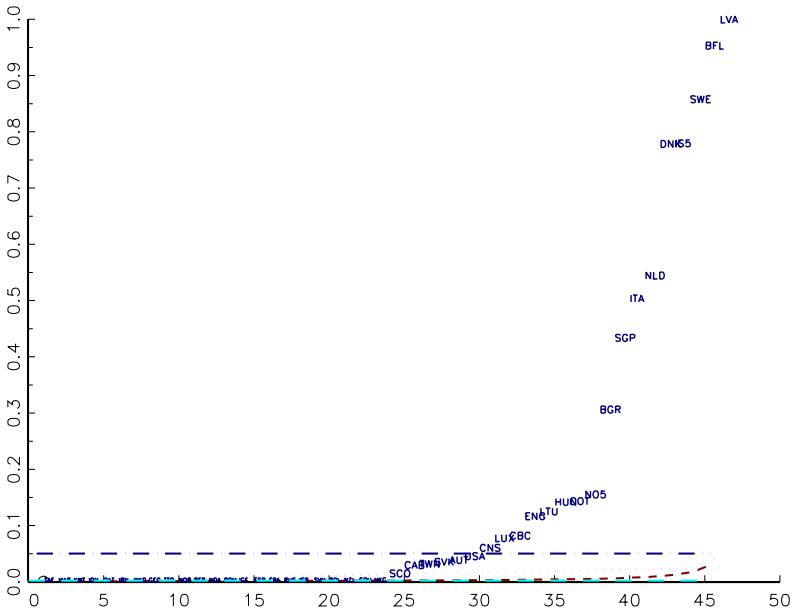
We first aggregate student achievement data at the school level. Due to missing data, the PIRLS database constructs imputation-based results for all students, so called "Plausible Values". The PIRLS reading achievement scale is standardized to have a mean of 500. All our results are based on "Plausible Value: Overall Reading PV5".[8] Then, let $T_i$ be the number of schools in country/region $i$; and $\bar{x}_{T,i}$ and $s^2_{T,i} = (n-1)^{-1} \sum_{t=1}^{T_i} (x_t - \bar{x}_{T,i})^2$ countrywise averages und variances across schools. We keep the statistical approach to the countrywise comparisons to Germany deliberately simple to focus on the effect of multiplicity on the test results.[9] Accordingly, we conduct a standard $t$-test, defined by the rejection of the null $H_0 : \mu = \mu_0$ if

$$|t_i| := |\sqrt{T_i}(\bar{x}_{T,i} - \mu_0)/s_{T,i}| > c_{\alpha/2},$$

[7]Of course, the choice of Germany is arbitrary and dictated by the author's personal interest. Any other country (as well as the overall average of 500) could have been used as a reference country.

[8]Since the different Plausible Values correlate rather strongly, all our results should remain qualitatively unchanged if another of the available imputed set of scores was used.

[9]Mullis et al. [2007] report standard errors based on jackknife estimates to take the stratified sampling design as well as imputation error into account. Again, differences to our results were checked to be negligible.



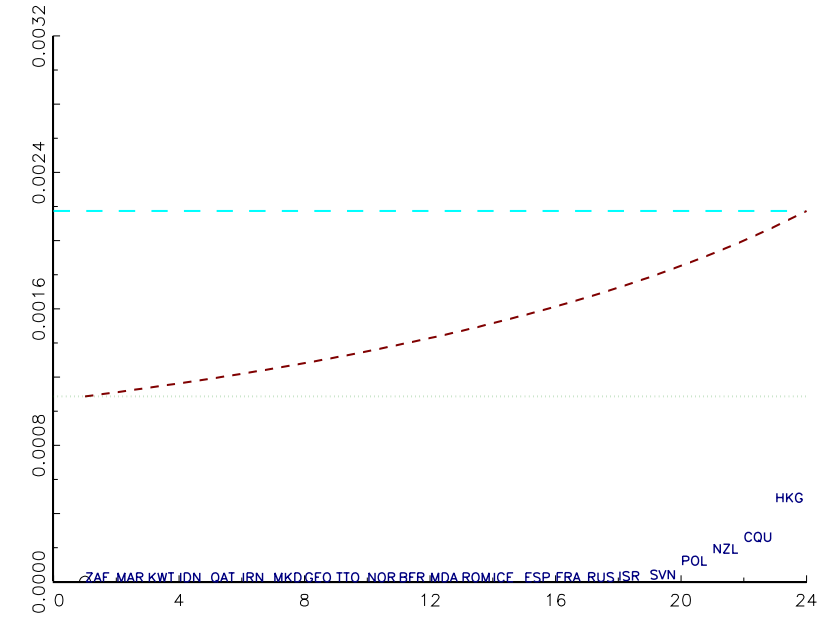

Blue Dotted Dashes—Standard Cyan Dashes—Hommel
Red Short Dashes—Holm Magenta Dots—BH

Figure III—Individual Countries' $p$-values
Relative to Germany

the $\alpha/2$-quantile of the normal distribution. The corresponding two-sided $p$-value is $p_i = 2(1 - \Phi(|t_i|))$, with $\Phi$ the standard normal distribution function. As Germany's 2006 average score is 548 (see Table II), we have $\mu_0 = 548$.

Results are presented in Figure III. (See Table II for the abbreviations used.) For a number of countries such as Sweden (SWE), Italy (ITA), the Netherlands (NLD) or England (ENG), the $p$-values are rather large, implying that student reading literacy in these countries' schools is by all standards comparable to that in Germany. For several countries, we find $p$-values less than, but close to, 0.05. Examples include Austria (AUT), the United States (USA) or Slovakia (SVK). Finally, there is a large set of countries for which the the $p$-values are essentially indistinguishable from zero. (To make the lower left portion of Figure III more easily readable, Figure IV reports that fraction in higher resolution.)

Several points are worth noting. Applying the 'standard' approach would lead to 29




Cyan Dashes—Hommel Red Short Dashes—Holm Green Dots—Bonferroni

Figure IV—Some Countries' $p$-values Relative to Germany

rejections, as $p_{(29)} = p_{\text{USA}} = 0.039$ and $p_{(30)} = p_{\text{CNS}} = 0.054$. On the other hand, all *FWER*-controlling techniques (cf. Sections 2.1-2.2) happen to indicate the same number of 23 rejections in this application. This is due to a large jump from $p_{(23)} = p_{\text{HKG}} = 0.0004$ to $p_{(24)} = p_{\text{SCO}} = 0.0085$, a value 20 times as large. This leads all these procedures to cut off at 23. Note, however, from Figure IV that the rejection curve of Hommel's procedure lies noticeably higher than that of Bonferroni's, and, initially, that of Holm's. Although, as seen above, the particular distribution of the $p$-values in this application implies that Hommel's procedure is no more rejective than Bonferroni's here, this underscores that the former is likely to be more powerful in general. According to the more liberal *FDR*-controlling Benjamini and Hochberg [1995]-procedure (with $\gamma = 0.025$) we find 24 rejections, as Scotland is also declared significantly different from Germany. All in all, the multiple testing procedures indicate 5-6 countries less for which significant differences in the ability of schools to foster reading literacy in primary school children relative

to Germany can be found. This implies that the number of different countries may be overstated by up to 30% by the 'standard' approach.

Of course, the two-sided $p$-values plotted in Figures III and IV give no indication as to whether schools in countries with small $p$-values are better or weaker than in Germany. Table III therefore reports the direction of a rejection according to the different (multiple) testing criteria. According to the 'standard' approach, more than half of the countries' schools (the entire left column plus those countries in the middle column printed in italics) are declared significantly weaker than Germany's. Three countries—those from the right column and Alberta are found to have significantly better schools. On the other hand, the multiple testing procedures paint a more cautious picture. Only 21 (22 in the case of the Benjamini and Hochberg [1995]-procedure) are found to have significantly weaker schools than Germany. Only two countries' schools—the Russian Federation and Hong Kong—are significantly more successful than Germany at conferring reading literacy to its students.

# 5 Conclusion

This study has investigated which countries' schools can be be classified as significantly better or weaker than Germany's as regards the reading literacy of primary school children. The 'standard' approach is to conduct separate tests for each country relative to the reference country (Germany) and to reject the null of equally good schools for all those countries whose $p$-value satisfies $p_i \leqslant 0.05$. It is discussed that this approach suffers from not controlling for multiplicity. That is, it overstates the difference between schooling systems by producing unwarranted rejections of the null. We demonstrate how various multiple testing techniques can remedy this problem. The results suggest that the 'standard' approach may overstate the number of significantly different countries by

Table III—Classification of Countries Reading Skills Relative to Germany.

| Weaker than Germany | As Good as Germany | Better than Germany |
|---|---|---|
| South Africa (ZAF) | *Scotland* (SCO) | Russian Federation (RUS) |
| Morocco (MAR) | Canada, Alberta (CAB) | Hong Kong (HKG) |
| Kuwait (KWT) | *Taiwan* (TWN) | |
| Indonesia (IDN) | *Slovakia* (SVK) | |
| Qatar (QAT) | *Austria* (AUT) | |
| Iran (IRN) | *USA* (USA) | |
| Macedonia (MKD) | Canada, Nova Scotia (CNS) | |
| Georgia (GEO) | Luxembourg (LUX) | |
| Trinidad & Tobago (TTO) | Canada, Brit. Columbia (CBC) | |
| Norway (NOR) | England (ENG) | |
| Belgium (French) (BFR) | Lithuania (LTU) | |
| Moldova (MDA) | Hungary (HUN) | |
| Romania (ROM) | Canada, Ontario (COT) | |
| Iceland (ICE) | Norway (5th grade) (NO5) | |
| Spain (ESP) | Bulgaria (BGR) | |
| France (FRA) | Singapore (SGP) | |
| Israel (ISR) | Italy (ITA) | |
| Slovenia (SVN) | Netherlands (NLD) | |
| Poland(POL) | Denmark (DNK) | |
| New Zealand (NZL) | Iceland (5th grade) (IS5) | |
| Canada, Quebec (CQU) | Sweden (SWE) | |
| (Scotland) (SCO) | Belgium (Dutch) (BFL) | |
| | Latvia (LVA) | |

Country classification according to the multiple testing procedures.
In brackets: Declared Significant by BH. In Italics: Declared weaker by the 'standard' approach.
In Smallcaps: Declared better by the 'standard' approach.

up to 30%.

Of course, the techniques employed here are by no means the only ones that could have been used. The multiple testing literature is very active in suggesting procedures that can be more suitable in related applications. Plausible candidates include resampling-based techniques [see, e.g. Romano and Wolf, 2005], or, in particular in applications with large $n$, further *FDR*-controlling procedures such as those of Finner, Dickhaus, and Roters [200x].

Furthermore, the approach used here could be extended to for instance efficiency analyses of countries' schools by comparing residuals from a suitable (e.g. panel) estimator relating reading achievement scores to input variables such as class size or investment per student.

We believe the framework put forward here may prove valuable in related applications. For instance, Jürges and Schneider [2007] propose a 'fair' ranking of teachers based on German PIRLS data (fair meaning that determinants of student achievements that are beyond the control of the teacher are controlled for). They then rank teachers into three different groups: average teachers and teachers that are 'significantly' better/weaker than the average, depending on whether suitable confidence intervals of a teacher's performance does (not) straddle the overall average efficiency. In view of the duality of tests and confidence intervals, such an approach will also suffer from "multiplicity", in the sense that a certain number of teachers will unduly be declared above or below average. When controlling for multiplicity, one would quite likely find that based on the available data, it would only be possible to declare fewer than Jürges and Schneider's 36.7% of teachers as truly different from the average.

Similarly, Wößmann and West [2006] conduct a multi-country study to investigate whether smaller classes lead to significant improvements in student achievement. They only find 'significantly' positive effects for Greece and Iceland. This finding is rationalized by noting that teachers in these countries are relatively less qualified and hence likely less well equipped to deal with large classes. It is beyond the scope of the present study to discuss whether the rationale put forward by Wößmann and West [2006] holds true—we do however tentatively suggest that these two significant findings might be induced by not controlling for multiplicity, rather than a genuine positive effect of smaller class sizes in these countries.

Similar issues arise in many studies in this and related literatures, see e.g. Hanushek and

Wößmann's [2006] study of the impact of early tracking on student achievement. As such, it could be a fruitful topic for further research to investigate whether these findings are robust to controlling for multiplicity, i.e. whether they are still significant according to the procedures such as those sketched in Section 2.

# References


Benjamini, Yoav, and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B* 57:1 (1995), 289–300.

Benjamini, Yoav, and Daniel Yekutieli, "The Control of the False Discovery Rate in Multiple Testing under Dependency," *The Annals of Statistics* 29:4 (2001), 1165–1188.

Bos, Wilfried, Sabine Hornberg, Karl-Heinz Arnold, Gabriele Faust, Lilian Fried, Eva-Maria Lankes, Knut Schwippert, and Renate Valtin (Eds.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (Münster: Waxmann, 2007).

Dudoit, Sandrine, and Mark J. van der Laan, *Multiple Testing Procedures and Applications to Genomics*, Springer Series in Statistics (Berlin: Springer, 2007).

Finner, Helmut, Thorsten Dickhaus, and Markus Roters, "On the False Discovery Rate and an Asymptotically Optimal Rejection Curve," *The Annals of Statistics* (to appear).

Hanushek, Eric A., and Ludger Wößmann, "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries," *The Economic Journal* 116 (2006), C63–C76.

Holm, Sture, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics* 6:1 (1979), 65–70.

Hommel, Gerhard, "A Stagewise Rejective Multiple Tests Procedure Based on a Modified Bonferroni Test," *Biometrika* 75:2 (1988), 383–386.

Jürges, Hendrik, and Kerstin Schneider, "Fair ranking of teachers," *Empirical Economics* 32 (2007), 411–431.

Lehmann, Erich L., and Joseph P. Romano, *Testing Statistical Hypotheses* (New York: Springer, 2005), 3rd ed.

Mullis, Ina V.S., Michael O. Martin, and Ann M. Kennedy Pierre Foy, *Progress in International Reading Literacy Study* (Boston: TIMSS & PIRLS International Study Center, 2007).

Romano, Joseph P., and Michael Wolf, "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica* 73:4 (2005), 1237–1282.

———, "Formalized Data Snooping Based on Generalized Error Rates," *Econometric Theory* 24 (2008), 404–447.

Sarkar, Sanat K., "Probability Inequalities for Ordered $MTP_2$ Random Variables: A Proof of the Simes Conjecture," *The Annals of Statistics* 26:2 (1998), 494–504.

Savin, N. Eugene, "Multiple Hypothesis Testing," in Z. Griliches, and M.D. Intriligator (Eds.), "Handbook of Econometrics," vol. 2, chap. 14 (Amsterdam: North-Holland Publishing, 1984), pp. 827–879.

Simes, R. John, "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika* 73:3 (1986), 751–754.

Wößmann, Ludger, and Martin West, "Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS," *European Economic Review* 50 (2006), 695–736.